Retail Sales Dataset Analytics

Jeisson Steve Rojas Velasquez

May 2025.

Introduction

This project aims to analyze a retail sales and customer demographics database. The goal is to identify customer consumption trends in retail products and the interaction between customers and products.

In addition, this database could help increase sales, attract more customers, and identify the best-selling product categories to improve the performance of retail stores.

Objectives

- Identify retail consumption trends based on gender and purchased products: beauty, clothing, and electronics.

-Identify, in a retail market database, which products are most frequently bought by customers: beauty, clothing, and electronics.

-Identify which products are consumed according to age.

Database

The database used for this project is:
https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset
This database is composed of:

1. **Transaction ID:** A unique identifier for each transaction, allowing tracking and reference.
2. **Date:** The date when the transaction occurred, providing insights into sales trends over time.
3. **Customer ID:** A unique identifier for each customer, enabling customer-centric analysis.
4. **Gender:** The gender of the customer (Male/Female), offering insights into gender-based purchasing patterns.
5. **Age:** The age of the customer, facilitating segmentation and exploration of age-related influences.
6. **Product Category:** The category of the purchased product (e.g., Electronics, Clothing, Beauty), helping understand product preferences.
7. **Quantity:** The number of units of the product purchased, contributing to insights on purchase volumes.
8. **Price per Unit:** The price of one unit of the product, aiding in calculations related to total spending.
9. **Total Amount:** The total monetary value of the transaction, showcasing the financial impact of each purchase.

With 1,000 records.

Methodology

An initial data cleaning was done using Power Query in Excel, removing duplicates and adjusting column formats. Then, SQL was used for specific calculations and data grouping, such as segmenting ages into groups. The data analysis and visualization were done using pivot tables in Excel. Interactive dashboards were also created in Power BI and Excel.

These visualizations helped to identify sales patterns, the most popular products, and opportunities to attract new customers based on age groups, giving useful tools to make data-driven decisions.

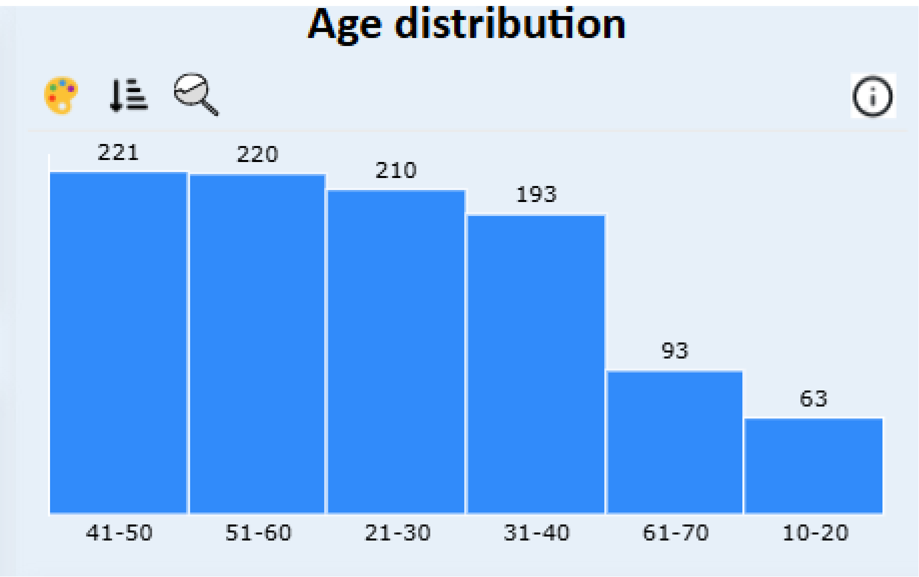
Descriptive statistics (based on the Total Amount column)

Measure	**Value**
Number of Observations (n)	1,000
Mean (Average)	$456.00
Median (Q2)	$135.00
Mode	Not applicable
Minimum	$25.00
Maximum	$2,000.00
1st Quartile (Q1)	$60.00
3rd Quartile (Q3)	$900.00
Interquartile Range (IQR)	$840.00
Variance (σ^2)	313,283.75
Standard Deviation (σ)	$559.72

Interpretation

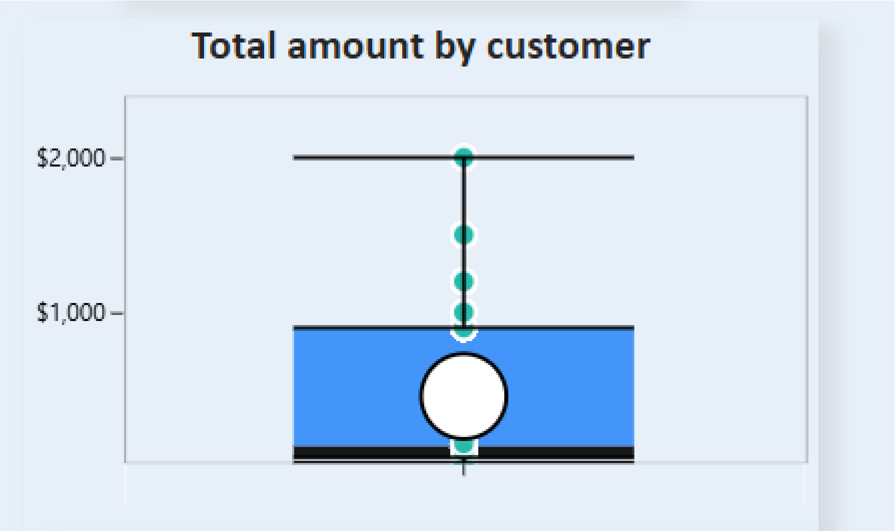
It is important to mention that there is a lot of variability in the amount of money spent by customers. The average ($456.00) is much higher than the median ($135.00), which shows that a few customers spend a lot and increase the overall sales average.

Charts

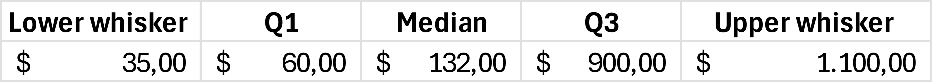


With this histogram, we can see a higher concentration of customers between 41 and 50 years old, representing around 22% of the total. On the other hand, customers between 61–70 and 10–20 years old had the lowest concentration.

This chart also shows a right-skewed distribution (positive skewness), which means the average is higher than the median.



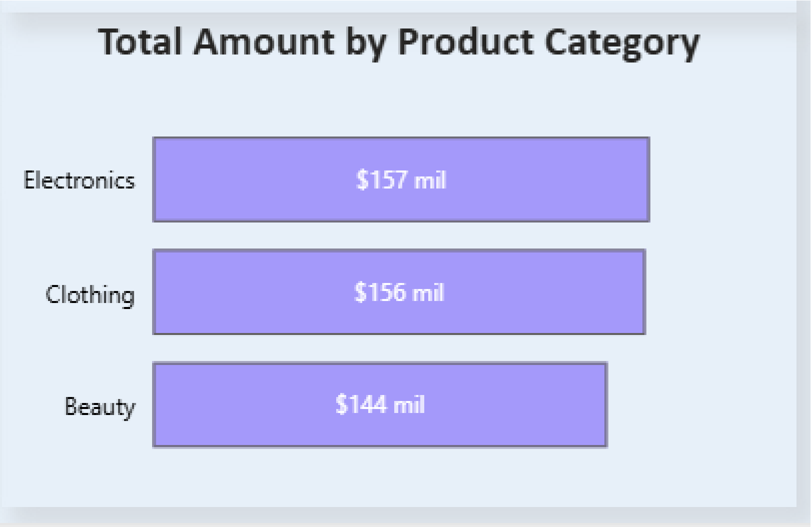
This is a boxplot, which was created using the following data:



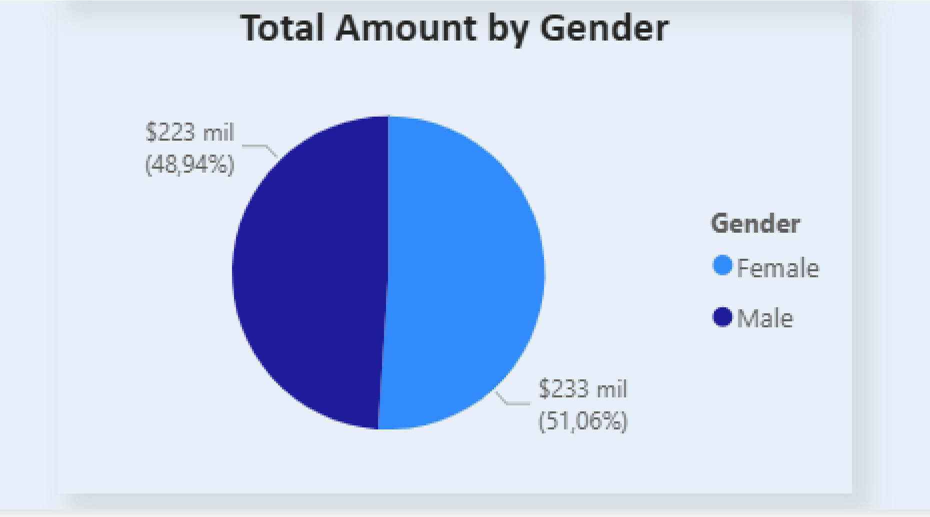
Lower whisker	Q1	Median	Q3	Upper whisker
$ 35,00	$ 60,00	$ 132,00	$ 900,00	$ 1.100,00

In this boxplot, we can see that 50% of the purchases (the interquartile range) are between $60 and $900 USD. The median, which shows the typical transaction value, is $132 USD. There are no outliers, because all the data points are inside the whiskers.

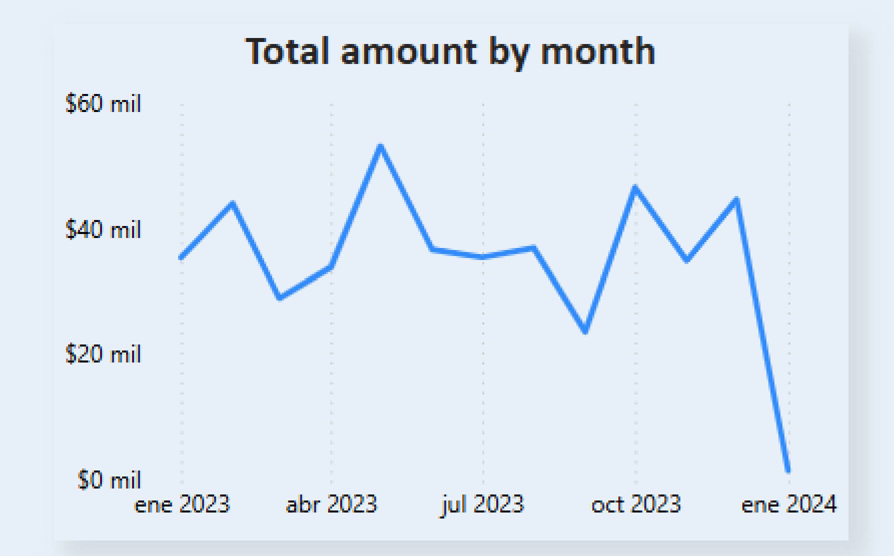
Also, there is a high dispersion, which means that the purchase amounts vary a lot and are not concentrated in a single range.



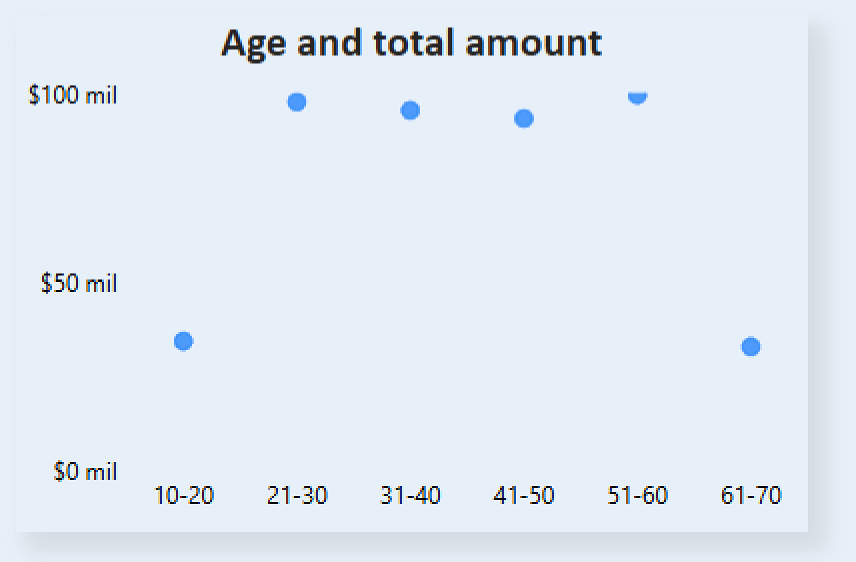
This bar chart shows the total revenue by product category. It highlights that the category with the highest revenue was Electronics, with a total of $157,000 USD, representing 34.2% of total sales. Even though fewer units were sold, it generated more income. On the other hand, the lowest revenue came from Beauty products, with a total of $144,000 USD.



This chart shows that there are more female buyers than male buyers (females represent 51.06%), which is more than half of the total buyers.

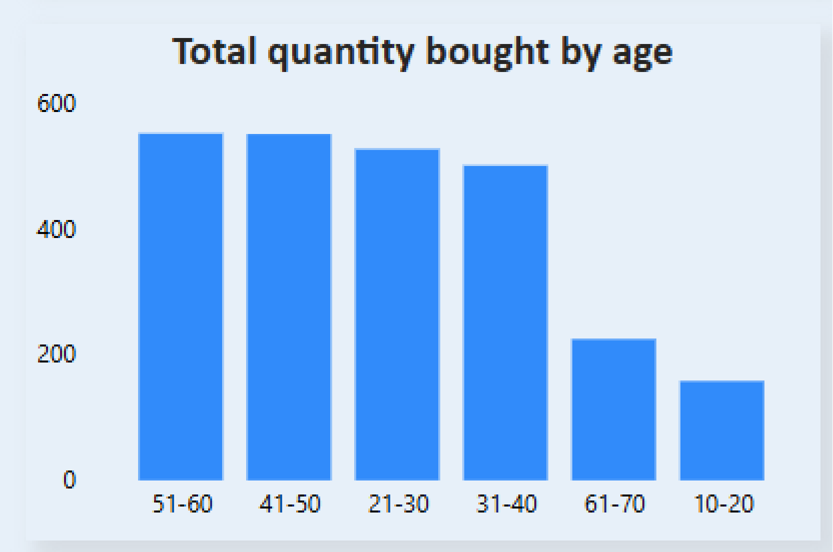


In this chart, there are significant peaks in purchases in February, May 2023, October, and December. However, overall, the sales performance during 2023 was not the best, with a noticeable drop from December to January 2024.

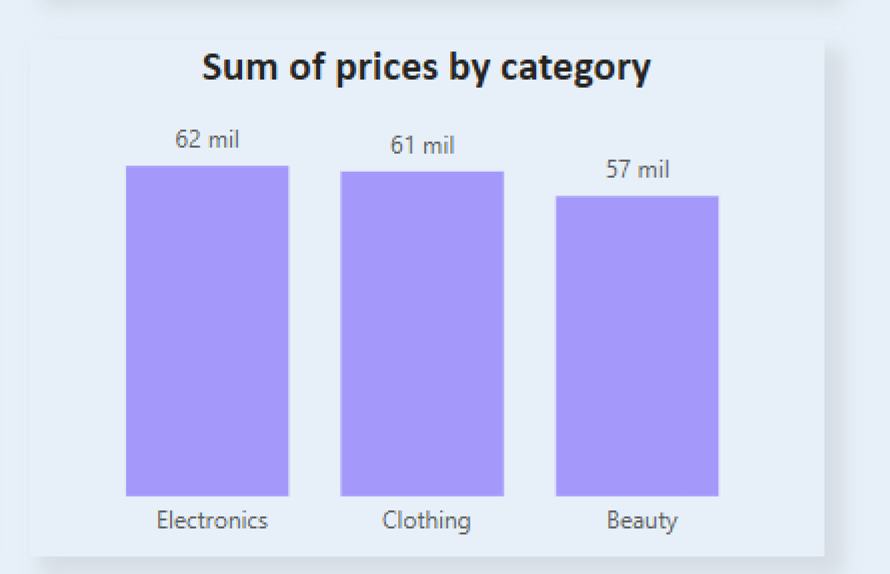


Regarding age, the age group 51-60 spent the most money, with a total of $100,085. This could be due to many factors, such as the possibility of being retired people or individuals with higher financial resources. Interestingly, people aged 61-70 spent less, and it's worth noting that very few people in this group made purchases.

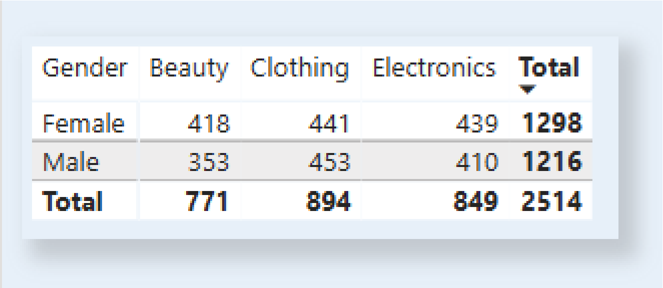
On the other hand, it is important to mention that the age ranges 21-30, with a total spending of $98,215; 31-40, with a total spending of $95,950; and 41-50, with a total spending of $93,795, show significant purchase amounts.



Also, this chart shows that the age group that buys the most products is 51-60, with a total of 552 purchases, representing 22% of the total. This chart also shows a kind of right skewness.

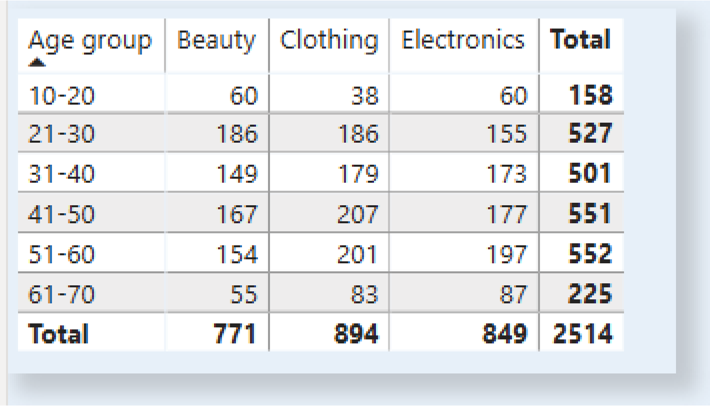


Regarding this chart, it shows that the product category with the highest total price sum of products sold was Electronics, with a total of $62,000 USD, representing 34.58% of the total price sum by product. This indicates that Electronics are the highest-priced products. Following this, Clothing had a total of $61,000 USD, with no significant difference, and finally, Beauty products had a total of $57,000 USD.



Gender	Beauty	Clothing	Electronics	**Total**
Female	418	441	439	**1298**
Male	353	453	410	**1216**
Total	**771**	**894**	**849**	**2514**

This table shows the quantities purchased by gender, highlighting that the largest purchases made by women were clothing, a trend also seen in men. Additionally, according to this chart, clothing had the highest sales compared to the other items.



Age group	Beauty	Clothing	Electronics	Total
10-20	60	38	60	**158**
21-30	186	186	155	**527**
31-40	149	179	173	**501**
41-50	167	207	177	**551**
51-60	154	201	197	**552**
61-70	55	83	87	**225**
Total	**771**	**894**	**849**	**2514**

Finally, this last table shows that, according to age ranges, people between 10-20 years old and 21-30 years old mostly bought clothing and electronics. Meanwhile, the age groups 31-40, 41-50, and 50-60 made their largest purchases in clothing, while those between 61-70 mostly bought electronics.

Storytelling

During the analysis of sales data from 2023 to 2024, we found that on average, each customer spent $456 USD per purchase. The highest value recorded in a single transaction was $2,000 USD, while the lowest spend was only $25 USD.

This large difference shows us that not all customers buy the same: some spend much more than others. In fact, the standard deviation was $559.72, indicating a high level of variability. This means that a few customers making expensive purchases are raising the average.

When looking at who is buying, we noticed that most buyers are between 41 and 50 years old. Additionally, half of the purchases (50%) were between $60 and $900, which gives us a better idea of the most common price ranges.

The "Electronics" category did not sell as many units but generated more money. The total price of the products sold in this category was $62,000 USD, showing that the products have high prices, while the total revenue from all sales in this category was $157,000 USD.

This highlights that, even though fewer electronics were sold, they generated more revenue.

We also found that 51.06% of the buyers were women, and their favorite product was clothing, with 441 items purchased. However, the age group 51-60 was the one that made the most purchases in total, regardless of the category.

Finally, sales were not uniform throughout the year. There was a significant drop between December and January, suggesting that after the high season at the end of the year, consumption decreased.

Findings

By analyzing the sales data, we can see that most buyers are in the age range of 51 to 60 years old, with a total of 221 people. This represents approximately 22% of the total buyers. This group made the most transactions, and their purchases were mainly concentrated in Clothing and Electronics products.

Regarding product categories, Electronics generated the most revenue, reaching a total of $157,000 USD, which is 34.2% of total sales. Additionally, when adding the unit prices of the products sold in this category, we get $62,000 USD, suggesting that the products have a high unit value. In summary, it was the most profitable category.

The month with the highest sales was May 2023, which we infer could be related to Mother's Day, as most of the buyers were women (51.06%), who also primarily bought clothing. On the other hand, there was a significant drop in sales between December 2023 and January 2024, which could have been caused by a seasonal dip after the holidays.

Overall, Clothing was the most sold category in terms of quantity, accounting for 35.1% of the total units sold, regardless of the buyers' age.

Finally, it is notable that women were the main buyers, representing just over half (51.06%) of the total, with a clear preference for clothing products.

Insights and recommendations

As recommendations, it is suggested to continue offering a wide variety of products for people aged 41-50, and to offer promotions on clothing and electronics, as well as a greater variety of products. It is also necessary to create offers for people in the 61-70 and 10-20 age ranges, and even consider the possibility of introducing new products for these age groups. It is recommended to conduct market research on consumption trends within these age ranges.

It is important to seek strategies that encourage a higher number of purchases, which can be driven by offers, discounts, promotions, and providing credit to encourage consumers to spend more, thus improving the company's revenue.

It is also important to include more electronics and even tech products due to their high demand. Additionally, the possibility of creating market strategies focused on beauty products should be evaluated, considering that most consumers in this product category are women.

Regarding the months, it is important to develop consumption events according to the season and special dates such as Mother's Day, Father's Day, Children's Day, Christmas, and New Year. As seen, May had a sales peak much higher than the other months. These types of strategies will improve monthly sales of products according to celebrations and special days.

Challenges

Data Limitations: It is important to mention that the samples of people aged 10-20 years were quite limited (only a total of 63) as well as participants in the 60+ age range (only a total of 93), which makes the analysis somewhat biased due to this.

Technical Difficulties: At this point, it is necessary to highlight the difficulty of recreating the Power BI dashboard in Excel. This was because some of the tables used to recreate graphs like the boxplot and histogram had to be copied and pasted as values only. This caused the tables not to be linked to the original database, preventing the use of time filters and slicers.

Results Interpretation: This was perhaps one of the biggest challenges, especially creating the boxplot in Excel and later interpreting it, which represented a real challenge. On the other hand, we could have worked on products sold and price per unit to get more detailed information and analyze it by age, but it was considered more general and broader to work with the total amount.